

On the Equivalence of Factorized Information Criterion Regularization and the Chinese Restaurant Process Prior

Shaohua Li

Nanyang Technological University, Singapore
shaohua@gmail.com

July 2, 2015

Abstract

Factorized Information Criterion (FIC) is a recently developed information criterion, based on which a novel model selection methodology, namely Factorized Asymptotic Bayesian (FAB) Inference, has been developed and successfully applied to various hierarchical Bayesian models. The Dirichlet Process (DP) prior, and one of its well known representations, the Chinese Restaurant Process (CRP), derive another line of model selection methods. FIC can be viewed as a prior distribution over the latent variable configurations. Under this view, we prove that when the parameter dimensionality $D_c = 2$, FIC is equivalent to CRP. We argue that when $D_c > 2$, FIC avoids an inherent problem of DP/CRP, i.e. the data likelihood will dominate the impact of the prior, and thus the model selection capability will weaken as D_c increases. However, FIC overestimates the data likelihood. As a result, FIC may be overly biased towards models with less components. We propose a natural generalization of FIC, which finds a middle ground between CRP and FIC, and may yield more accurate model selection results than FIC.

1 Equivalence of FIC and CRP when $D_c = 2$

Suppose there are a sequence of 1-of- K latent coding variables $\mathbf{Z} = \mathbf{z}_1, \dots, \mathbf{z}_N$. For any k , let $n_k = \sum_{i=1}^N z_{ik}$. Then \mathbf{Z} corresponds to a partition of N numbers into K sets S_1, \dots, S_K , where $|S_k| = n_k$. This partition is denoted as $\mathbf{B} = (S_1, \dots, S_K)$. The correspondence between \mathbf{Z} and the partition \mathbf{B} is referred to as \mathbf{Z} maps to \mathbf{B} , denoted as $\mathbf{Z} \mapsto \mathbf{B}$.

A Chinese restaurant process [3, 2] assigns to this sequence a prior probability

$$P_{\text{CRP}}(\mathbf{Z}) = \frac{\prod_k (n_k - 1)!}{N!K!}.$$

There are $\frac{N!}{\prod n_k!}$ configurations of \mathbf{Z} mapping to the same \mathbf{B} . These configurations of \mathbf{Z} form an equivalence class $\{\mathbf{Z}|\mathbf{Z} \mapsto \mathbf{B}\}$. When it is clear from context, we also denote $\mathbf{B} = \{\mathbf{Z}|\mathbf{Z} \mapsto \mathbf{B}\}$. The probability of this equivalence class is:

$$P_{\text{CRP}}(\mathbf{B}) = \frac{N!}{\prod n_k!} P(\mathbf{Z}_0) = \frac{1}{K! \prod_{n_k > 0} n_k}, \quad (1)$$

where \mathbf{Z}_0 is any configuration that maps to \mathbf{B} .

Note that K is a free parameter. Fixing K to a particular value, we obtain a distribution of \mathbf{B} conditioned on K :

$$P_{\text{CRP}}(\mathbf{B}|K) = \frac{1}{\mathcal{Z}_K} \frac{1}{\prod_{n_k > 0} n_k},$$

where $\mathcal{Z}_K = \sum_{\sum n_k = N} \frac{1}{\prod_{n_k > 0} n_k}$ is the normalizing constant.

When $D_c = 2$, the FIC regularization term in [1, eq.9] is

$$P_{\text{FIC}}(\mathbf{Z}|K) \sim \frac{1}{\prod_{n_k > 0} (\sum z_{ik})} = \frac{1}{\prod_{n_k > 0} n_k}. \quad (2)$$

By comparing (1) and (2), one can see that this regularizer term is equivalent to the CRP prior over the equivalence class when the model parameter dimensionality $D_c = 2$.

2 Stronger Model Selection of FIC when $D_c > 2$

In higher dimensionality of D_c , $P_{\text{FIC}}(\mathbf{Z}|K) \sim \frac{1}{\prod_{n_k > 0} n_k^{D_c/2}}$, i.e. the FIC regularizer becomes sharper and more biased among different configurations of \mathbf{Z} . To analyze the significance of the exponent $D_c/2$, suppose we use a prior $p(\mathbf{Z})$ of \mathbf{Z} in a model, where the data likelihood is given by $p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta})$. The posterior of \mathbf{Z} is proportional to $p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta})p(\mathbf{Z})$.

We first suppose the configuration of \mathbf{Z} is known, and consider a Laplace approximation of $p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$. When D_c is larger, $p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta})$ decreases more quickly when $\boldsymbol{\theta}$ deviates from the ML estimator (MLE) $\bar{\boldsymbol{\theta}}$. That is, in $p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta})p(\mathbf{Z})$, if D_c is large enough, the effect of $p(\mathbf{Z})$ will be dominated by $p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta})$, if the prior $p(\mathbf{Z})$ does not change with D_c . In other words, the regularization brought about by the prior will weaken as D_c increases. The CRP prior is covered by this analysis. In contrast, as the FIC regularizer becomes sharper as D_c increases, the domination of $p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta})$ over $p(\mathbf{Z})$ will not happen with the increase of dimensionality of $\boldsymbol{\theta}$. This suggests that FIC will have a stronger model selection effect and tend to be more ‘‘parsimonious’’ than CRP, when the parameter dimensionality is high.

3 Possible Limitations and Generalizations of FIC

The analysis in Section 2 is based on Laplace approximation, in which the approximating Gaussian is only accurate around a small area around the MLE

$\bar{\theta}$ of $p(\mathbf{X}|\mathbf{Z}, \theta)$. It might be the case that the estimated marginal probability is greater than the actual probability. Actually during the derivation of Factorized Asymptotic Bayesian inference, [1] assumes that the MLE $\bar{\theta}$ of $\sum_{\mathbf{Z}} p(\mathbf{X}|\mathbf{Z}, \theta)p(\mathbf{Z})$ is also the MLE of $p(\mathbf{X}|\mathbf{Z}, \theta)$ for each particular \mathbf{Z} . Therefore the estimated marginal probability would be greater than the actual marginal probability.

In this regard, we could extend FIC to a Generalized FIC (GFIC), which is milder thanks to a smaller exponent, i.e. $P_{\text{GFIC}}(\mathbf{Z}|K) \sim \frac{1}{\prod_{n_k > 0} n_k^d}$, where $1 < d < D_c/2$.

References

- [1] R. Fujimaki and S. Morinaga. Factorized asymptotic bayesian inference for mixture modeling. In *AISTATS*, volume 22, pages 400–408, 2012.
- [2] Samuel J Gershman and David M Blei. A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12, 2012.
- [3] Carl Edward Rasmussen. The infinite gaussian mixture model. In *NIPS*, volume 12, pages 554–560, 1999.